POHON KEPUTUSAN DALAM PENGKLASIFIKASIAN PENJURUSAN SISWA SEKOLAH MENENGAH ATAS (SMA)

Amalia Anjani Arifiyanti[1], Anisa Lucky Ana[2], dan Ayu Dwi S.[3]
Jurusan Sistem Informasi, Institut Teknologi Adhi Tama Surabaya[1,2,3]
anjani.arifiyanti@itats.ac.id

ABSTRACT

Determining the majors of high school studentd are based on several predetermined criteria. The criteria are score/grade of student on all subjects in the first year of high school. The use of data mining classification using decision tree C4.5 algorithm is expected to assist teacher in accelerating the process of decission and to minimize the risk for determining the students' major. The accuracy of this study resulted 94.6%, average precision is 0,951, and average recall is 0,946 in classification of high school students into two majors that are science major and social major.

Kata kunci: *C45, Decision Tree, Classification, Senior High School*

ABSTRAK

Penjurusan siswa Sekolah Menengah Atas (SMA) hingga saat ini ditentukan berdasarkan beberapa kriteria yang telah ditentukan. Kriteria tersebut yakni nilai siswa pada seluruh mata pelajaran pada saat siswa tersebut berada pada tahun pertama SMA. Penggunaan klasifikasi data mining dengan menggunakan algoritma pohon keputusan C4.5 diharapkan mampu membantu para guru SMA dalam mempercepat proses penjurusan para siswa tersebut dan meminimalisir kesalahan penjurusan yang mungkin akan terjadi. Model klasifikasi yang dihasilkan memiliki tingkat akurasi sebesar 94.595%, rata-rata precission sebesar 0,951 dan rata-rata recall sebesar 0,946 dalam pengklasifikasian para siswa SMA ke dalam dua jurusan yakni sains dan sosial.

Kata kunci: *C45, Decision Tree, Klasifikasi, Sekolah Menengah Atas*

PENDAHULUAN

Pelaksanaan wajib belajar 12 tahun[1] mendorong masyarakat untuk mengikuti pendidikan formal dari sekolah dasar (SD) hingga Sekolah Menengah Atas (SMA). Peserta jenjang pendidikan formal terakhir yaitu Sekolah Menengah Atas (SMA) secara umum berusia 15-18 tahun yang mengikuti proses pembelajaran mulai dari kelas 10 hingga kelas 12. Pada tahun pertama atau kelas 10, siswa SMA mendapatkan berbagai mata pelajaran umum. Tetapi pada tahun kedua atau kelas 11, siswa SMA diwajibkan memilih salah satu dari tiga jurusan yang ada yakni sains, sosial, dan bahasa walaupun pada beberapa sekolah hanya terdapat dua jurusan saja yaitu sains dan sosial. Pada kelas 11 dan 12 para siswa akan mendapat mata pelajaran yang mengikuti kurikulum masing-masing jurusan[2]. Penjurusan siswa ini dilakukan untuk menggali potensi dalam diri siswa, minat, dan bakat yang dimiliki sehingga hal ini akan dapat membantu mengoptimalkan kemampuan masing-masing siswa. Penjurusan ini berdasarkan penilaian mata pelajaran yang diikuti para siswa pada kelas 10. Proses pengambilan keputusan dalam penjurusan masing-masing siswa ini dilakukan oleh para guru di kelas 10. Proses pengambilan keputusan ini akan beresiko tinggi jika jumlah data siswa di sekolah bertambah atau jumlah siswa cukup banyak. Resiko yang muncul adalah penjurusan siswa yang kurang tepat. Resiko pengambilan keputusan yang kurang tepat dapat mengakibatkan kerugian terhadap siswa. Kerugian tersebut diantaranya adalah siswa tertekan dalam proses pembelajaran hingga kurang tergalinya potensi siswa tersebut. Hal ini juga akan berpengaruh dalam pemilihan bidang ilmu bagi siswa yang akan melanjutkan ke perguruan tinggi.

Resiko kesalahan penjurusan siswa SMA tersebut dapat diminimalisir melalui metode klasifikasi dalam penggalian data. Klasifikasi merupakan proses menempatkan suatu objek ke

dalam satu kelompok kategori berdasarkan keseuaian karakteristik objek yang bersangkutan dengan target kelompok [3]. Tujuan klasifikasi adalah untuk membuat model dan menggunakannya untuk memprediksi kelas suatu obyek yang kelasnya belum diketahui [4]. Berdasarkan penelitian yang dilakukan oleh Algoritma C4.5 yang merupakan salah satu jenis pohon keputusan dipilih sebagai algoritma pengklasifikasi jurusan siswa SMA. Diharapkan klasifikasi dalam meminimalisir resiko salah penjurusan yang terjadi di SMA.

TINJAUAN PUSTAKA

Pohon Keputusan (*Decision Tree*)

Metode pohon keputusan digunakan secara luas dalam metode klasifikasi penggalian data. Pohon keputusan menggunakan representasi struktur pohon dari akar, cabang, hingga daun. Setiap atribut direpresentasikan dalam node dan node yang paling atas disebut dengan *root*, nilai dari atribut direpresentasikan sebagai cabang, dan kelas direpresentasikan oleh daun [5]. Setiap node cabang merupakan pilihan beberapa alternatif dan daun merupakan keputusan. Pohon keputusan dapat diartikan dalam bentuk aturan/skenario. Pohon keputusan dibentuk dengan berbagai jenis algoritma antara lain ID3, CART, dan C4.5.

Pohon keputusan mampu menangani data dengan skala yang berbeda, distribusi kelas yang tidak merata, fleksibel, dan mampu menangani hubungan linear antara fitur/atribut dan kelas [6]. Pohon keputusan dalam dilatih dengan cepat dan cepat dalam proses eksekusi [7]. Berbeda dengan algoritma pengklasifikasi lainnya yang proses klasifikasinya lebih ke arah *'black box'*, para analis dapat dengan mudah mengintepretasikan model yang dihasilkan pohon keputusan. Pohon keputusan juga dapat menangani data-data yang hilang atau bahkan *noise* [8].

Algoritma C45

Salah satu algoritma pohon keputusan adalah C4.5. Algoritma ini digunakan untuk membentuk model klasifikasi dalam bentuk pohon keputusan [9]. Secara umum tahapan pembentukan pohon keputusan menggunakan C4.5 adalah sebagai berikut [10]:

1. Menyiapkan data latih. Data latih diambil dari data terdahulu yang telah memiliki kelas sehingga dapat menjadi acuan bagi prediksi kelas pada data uji.
2. Pilih atribut sebagai *node*. Pemilihan atribut ini ditentukan melalui penghitungan nilai gain dari masing-masing atribut. Atribut yang memiliki nilai *Gain* paling tinggi akan menjadi *node*. Berikut adalah rumus *Gain*:

$$\Delta = I(parent) - \sum_{j=1} \frac{N(v_j)}{N} I(v_j) \quad \text{..... (1)}$$

Dengan:

I (*parent*) = nilai ketidak murnian *parent* node (direpresentasikan dengan nilai entropy dari parent node)
I (Vj) = nilai ketidakmurnian *child* node
N = jumlah *record* dalam *parent* node
N (Vj) = jumlah *record* dalam *child* node

Nilai entropy dihitung terlebih dahulu untuk mendapatkan nilai gain, nilai entropy dihitung dengan rumus sebagai berikut:

$$Entropy\ (t) = -\sum_{i=0}^{c-1} p(i|t)\ log_2\ p(i|t) \quad \text{..... (2)}$$

Dengan:

p = pecahan
i = jumlah *record* pada kelas i
t = jumlah *record* dalam node

3. Buat cabang untuk setiap node.
4. Ulangi tahap 2 dan 3 hingga:
   a. Seluruh *record* pada *node* masuk dalam kelas yang sama atau entropy bernilai 0
   b. Seluruh atribut telah digunakan
   c. Node kosong atau tidak memiliki *record.*

Evaluasi dan Validasi Model Klasifikasi

Model klasifikasi yang dihasilkan setelah memproses data latih dengan algoritma pengklasifikasi perlu dievaluasi untuk mengetahui performa model klasifikasi dalam memprediksi kelas. Metode evaluasi klasifikasi menggunakan metode confusion matrix [3]. *Confusion matrix* dapat dilihat pada tabel 1 berikut ini.

Tabel 1. *Confusion Matrix*

		Predicted Class	
		Class 0	Class 1
Actual Class	Class 0	TP	FN
	Class 1	FP	TN

Dengan :
TP : *True Positive* TN : *True Negative*
FN : *False Negative* FP : *False Positive*

Dari *confusion matrix* dapat diketahui nilai performa dari model klasifikasi, berikut rumus perhitungan akurasi, *precision*, dan *recall*:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (3)}$$

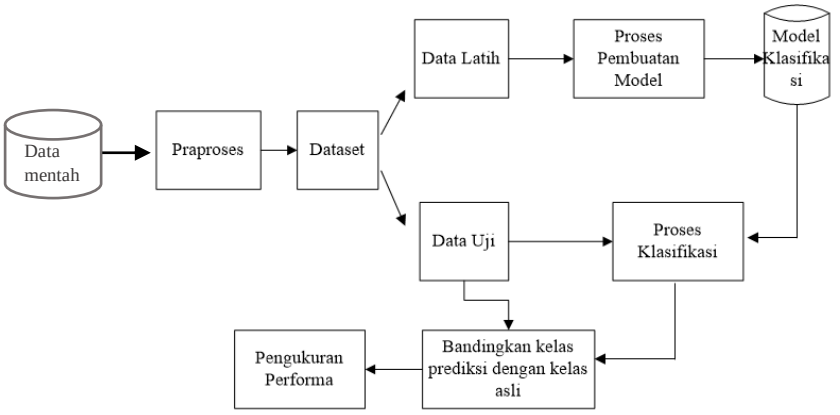
$$Precision = \frac{TP}{TP+FP} \quad \text{..... (4)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{..... (5)}$$

Akurasi merujuk pada tingkat keberhasilan model klasifikasi dalam melakukan keseluruhan prediksi kelas dengan benar. *Precission* didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. *Precision* merupakan probabilitas bahwa sebuah item yang dipilih adalah relevan. *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. *Recall* merupakan probabilitas bahwa suatu item yang relevan akan dipilih. Untuk precision dan recall dihitung per kelasnya, sehingga untuk mengetahui precision dan recall model klasifikasi dihitung rata-rata dari precision dan recall masing-masing kelas.

METODE

Untuk memprediksi jurusan bagi masing-masing siswa SMA dilakukan beberapa tahap klasifikasi. Secara umum tahap klasifikasi dapat dilihat pada gambar 1.



Gambar 1. Tahap Klasifikasi

Data yang digunakan pada penelitian ini adalah data nilai para siswa pada kelas 10 di SMA Negeri 1 Driyorejo, Gresik. Total data yang digunakan sebanyak 37 *record*. Atribut yang digunakan sebagai prediksi sebanyak tujuh atribut yaitu nilai mata pelajaran matematika, kimia, fisika, biologi, sejarah, geologi, dan sosial. Satu atribut kelas yang terdiri dari dua kelas yaitu sains dan sosial. Tahap praproses digunakan untuk menyesuaikan jenis data dengan kebutuhan. Data nilai dari masing-masing mata pelajaran dikonversi menjadi 'baik' dan 'sangat baik'. Jenis kelas yang digunakan adalah dua jenis yaitu sains dan sosial dengan banyaknya kelas sains adalah 17 *record* dan sosial adalah 20 *record*. Contoh data hasil praproses dapat dilihat pada tabel 1 sebagai berikut.

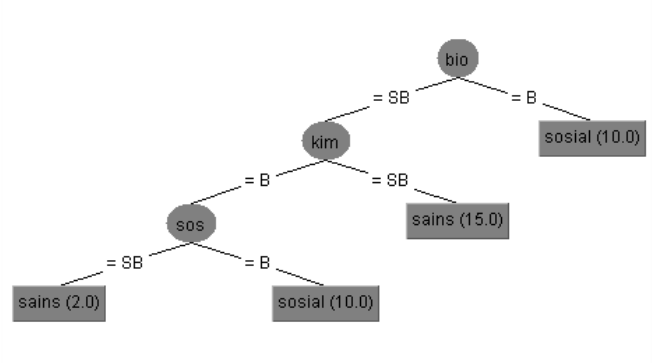
Tabel 2. Contoh *Dataset*

ID	Mat	Kim	Fis	Bio	Sej	Geo	Sos	Jurusan
1	SB	B	SB	SB	B	B	SB	Sains
2	SB	B	SB	SB	B	B	SB	Sains
3	SB	B	SB	B	B	B	SB	Sosial
4	SB	SB	SB	SB	B	B	B	Sains
5	SB	SB	SB	B	B	B	B	Sosial

Data yang telah diolah tersebut dibagi menjadi dua jenis data yaitu data latih dan data uji. Pembagian data latih dan data uji menggunakan metode *cross-validation* 10-*folds*. Metode pembagian data ini digunakan dikarenakan terbatasnya jumlah dataset. Model klasifikasi dibuat berdasarkan data latih. Pembuatan model klasifikasi ini menggunakan algoritma C4.5 yang prosesnya telah dijelaskan pada bagian sebelumnya. Model klasifikasi dalam bentuk pohon keputusan digunakan sebagai dasar untuk melakukan prediksi kelas pada data uji. Hasil prediksi kelas pada data uji akan dibandingkan dengan kelas sebenarnya. Tahap evaluasi ini dituliskan dalam bentuk *confusion matrix*. Perhitungan performa dari model klasifikasi didapatkan dari perhitungan akurasi dari perbandingan kelas prediksi dengan kelas sebenarnya.

HASIL DAN PEMBAHASAN

Model klasifikasi dalam bentuk pohon keputusan yang dibuat dari data latih dan menjadi dasar untuk proses prediksi pada data uji dapat dilihat pada gambar 2 berikut ini.



Gambar 2. Model Klasifikasi dalam Bentuk Pohon Keputusan

Pada model tersebut diketahui hanya tiga atribut yang digunakan yaitu biologi, kimia, dan sosial. Hal ini dimungkinkan karena terbatasnya data latih dan berdasarkan data latih tersebut atribut yang memiliki *gain* tinggi atau sebagai atribut prediktor yang kuat adalah tiga atribut tersebut sedangkan 4 atribut lainnya yaitu matematika, fisika, sejarah, dan geologi tidak dianggap sebagai prediktor yang cukup kuat.

Sedangkan hasil performa dari model klasifikasi tersebut dapat dilihat pada *confusion-matrix* berikut ini.

Tabel 3. *Confusion-Matrix Model Klasifikasi*

		Kelas prediksi	
		sains	sosial
Kelas sebenarnya	sains	15	2
	sosial	0	20

Dari *confusion-matrix* tersebut, diketahui bahwa 35 *record* diklasifikasikan dengan benar yaitu kelas prediksi sesuai dengan kelas prediksi tetapi terdapat dua *record* yang seharusnya diklasifikasikan sebagai kelas sains namun diprediksikan sebagai kelas sosial. Kesalahan klasifikasi tersebut yang menyebabkan model klasifikasi memiliki tingkat *error* sebesar 5.405 %, akan tetapi karena model klasifikasi tersebut memiliki tingkat akurasi yang cukup baik yakni sebesar 94.595%. Model klasifikasi tersebut memiliki tingkat rata-rata *precision* sebesar 0,951 dan rata-rata *recall* sebesar 0,946. Kesalahan prediksi dalam klasifikasi tersebut terjadi karena model klasifikasi yang belum mampu mencakup keseluruhan kondisi yang muncul. Atribut prediksi yang digunakan dalam model klasifikasi hanya tiga atribut yaitu biologi, kima, dan sosial sedangkan empat atribut lainnya tidak digunakan. *Records* yang salah kelas dapat dilihat pada tabel 4 berikut ini.

Tabel 4. Salah Prediksi

ID	Mat	Kim	Fis	Bio	Sej	Geo	Sos	Asli	Prediksi
1	SB	B	SB	SB	B	B	SB	Sains	Sosial
2	SB	B	SB	SB	B	B	SB	Sains	Sosial

KESIMPULAN

Pohon keputusan yang dibuat menggunakan algoritma C4.5 dapat digunakan dengan baik untuk memprediksikan jurusan para siswa SMA kelas 10. Model klasifikasi yang dihasilkan hanya menggunakan tiga atribut prediktor dari tujuh atribut yang digunakan. Tidak digunakannya seluruh

atribut dimungkinkan karena jumlah data latih yang digunakan terbatas yaitu 37 *record*. Namun demikian tingkat akurasi model klasifikasi yang dihasilkan cukup baik yaitu sebesar 94.595% dengan tingkat *error* sebesar 5.405 %. Tingkat rata-rata *precision* dan *recall* secara berurutan adalah 0,951 dan 0,946. Sehingga dapat dinyatakan bahwa klasifikasi dengan menggunakan algoritma C4.5 dapat digunakan sebagai salah satu alternatif untuk membantu para guru untuk memutuskan penjurusan siswanya pada kelas 11 dan 12.

DAFTAR PUSTAKA

[1] P. R. Indonesia, "Instruksi Presiden Republik Indonesia Nomor 7 Tahun 2014 tentang Pelaksanaan Program Simpanan Keluarga Sejahtera, Program Indonesia Pintar, dan Program Indonesia Sehat untuk Membangun Keluarga Produktif," 2014.

[2] A. Mangkoesapoetra, "Memberdayakan MGMP, Sebuah Keniscayaan," [Online]. Available: http://re-searchengines.com/art05-14.html. [Diakses 2 Maret 2016].

[3] F. Gorunescu, Data Mining Concept Model and Techniques, Berlin: Springer, 2011.

[4] J. &. K. M. Han, Data Mining Concept and Tehniques, San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3, 2006.

[5] D. T. Larose, Discovering Knowledge in Data, New Jersey: John Willey & Sons, Inc. ISBN 0-471-66657-2, 2005.

[6] M. A. &. B. C. E. Friedl, "Decision tree classification of land cover from remotely sensed data," *Remote Sensing of Environment*, no. 61, p. 399– 409., 1997.

[7] M. &. G. W. Gahegan, "The classification of complex geographic datasets: An operational comparison of artificial neural network and decision tree classifiers," dalam *Third International Conference on GeoComputation*, 1998.

[8] S. &. X. D. Dua, Data Mining and Machine Learning in Cybersecurity, USA: Taylor & Francis Group. ISBN-13: 978-1-4398-3943-0, 2011.

[9] S. P. Utari, "Implementasi Metode C4.5 untuk Menentukan Guru Terbaik," *Pelita Informatika Budi Darma*, vol. 9, no. 3, pp. 2301-9425, 2015.

[10] M. S. V. K. Pang-ning Tan, Introduction to Data Mining, Pearson, 2005.